1. Before we get serious ...
    1. Comparison - The Problem
    2. Comparison - A Dawning Light
    3. Comparison - A Retrospection
    4. Behind the Scenes
2. Overview
    1. Wordcorr overview
    2. Wordcorr tools
3. Metadata
    1. The Wordcorr Community
        1. Information about you
    2. Collections
        1. The shape of collections
        2. Collection metadata
        3. Collection titles
        4. Collection creator and collaborators
        5. Collection glosses
        6. Collection description
        7. Collection coverage
        8. Collection published sources
        9. Collection permanence
    3. Varieties
        1. The shape of speech varieties
        2. Variety names and codes
        3. Variety locations
        4. Variety sources
4. Language data
    1. The Data panel
    2. Wordcorr entries
    3. Type & edit
        1. Data from the keyboard

    2. Data editing
    3. Data with phonetic characters using Easy IPA
4. Backup and Exports
    1. Exporting for backup
    2. Exporting part of a collection
    3. Restoring from backup
5. Imports
    1. Import from XML
    2. Import from Elsewhere
5. All the Rest
    1. What We've Covered
    2. What We Haven't Covered
6. Appendix: Set up
    1. Set up Wordcorr
    2. First-time installation
    3. Upgrading Wordcorr

Comparison - The Problem

“Watson! I hoped you'd come soon!” Mr Sherlock Holmes greeted me as I came in the door, late in a Tuesday afternoon, October 9, 1888.

I quickly replied, “I dropped in at the Science Museum, which now houses the old Patent Office Museum collections. Charles Babbage's plan for an analytical engine caught my eye. I'm afraid I lost all track of time, talking to the person who is putting together a working model. I also read about the efforts a generation ago of the Countess of Lovelace to direct its working by indicating the order in which its available operations are to proceed. Babbage's device could do much for us physicians!”

“A perfect logical engine indeed, after my own heart!” exclaimed Holmes. “I must have a look myself some time. But not now: I have a telegram from the Foreign Office that they seek consultation on a matter. You'll stay, won't you?”

Before I could answer we heard a horse and cab stop at the kerb, followed by a pull on the bell. Mrs Hudson brought the caller up.

“Mr Sherlock Holmes, I presume? My name is Bond -- Rupert Bond, on Her Majesty's service.”

“Pray tell, Mr Bond, what brings you here. You may speak as confidently in front of Dr Watson as if you spoke to me alone. You are, I see by the healthy yet darkish tone of your skin, but newly returned from the tropics.”

“From the Malay States, sir, whence I return next week. As you may know, the British East India Company has developed its enterprises there so well, that even the minor rajahs are now linking up with the Company in hopes of sharing in its prosperity.

“The Company has been getting feelers from yet farther east, where Portugal dominates the Spice Islands and Holland would like to. And coming back to our English boiled mutton after savouring the fine cooking of the Malay States, I fully concur with Government's interest in extending our trade beyond its present limits.”

“But what,” injected Holmes, “brings you to me?”

Bond replied, “The Indies are thousands of islands, separated by dangerous channels. One trade route for us might go from Singapore to British North Borneo, on to the Straits of Makassar, thence south, then east, to the trading centers for spices. Less infested by pirates and Dutchmen than a straight run through the Java Sea, you know.

“For a time, the Kingdom of Makassar on the southwest of the island of Celebes welcomed free trade. But over two centuries ago the freebooters of Bugis, across the peninsula from Makassar, joined the Dutch in taking over that unfortunate city. Now the Dutch control those straits.

“South of Celebes lies Saleyer Island, a narrow piece of land separated by a deep and treacherous channel. It has long been a stopoff both for east-west and north-south travel. Our current interests lie to the east of there: Saleyer could be our stepping stone to the spice trade – if its people were to become convinced that we and Makassar are better partners than the Dutch and their allies.

“To ascertain how strong the bond is among these three peoples, my superiors have urged me to look into the methods for comparing languages pioneered by a judge in the British Raj, over a century ago. If the language of Saleyer proves more like that of Makassar than like Buginese, could we not suppose a greater affinity among those peoples, that could open a door for us?”

“But why,” rejoined Holmes, “why do you not rather ask this of a philologist, than of a detective?”

“Because, sir, the academic mind, be it ever so finely tuned, is infernally slow about arriving at a conclusion; whereas you, sir, are renowned for digesting facts quickly and coming up with answers! Nor do I see that their use of logic is all that different from what I read about yours,” he said with a nod toward me. “I sail next week.”

Holmes smiled slightly in agreement, and the other pulled from his pocket some sheets of paper. On them were written lists of words he had taken

down from the three languages. The first sheet began like this:

Makassar (M), Bugis (B), Saleyer (S)

	M	B	S
1. tree	poko'	poong	po'ong
2. leaf	leko'	daung	taha
3. root	aka'	uRi'	aka'
4. wood	----	'aju	kaju
5. thorn	katinting	duri	katinting
6. husk	ka'daro	ka'daro	ka'daro
7. coconut	kaluku	kaluku	anjoro
8. banana	unti	utti	looka
9. betelnut	leko	ota	koomba
10. rattan	Raukang	Raukang	lauro
11. bamboo	bulaw	piring	paRing
12. flower	bunga	bunga	bunga
13. fruit	Raponna	buwa	buwa
14. grass (?)	Rea	area	Rea

From the Bond papers

“I regret the wear and tear that render these papers somewhat hard to read, but if you must know, I had to leave the area rather in haste,” said our guest. “I also realize that I do not know how to write foreign speech sounds, nor am I conversant with the International Phonetic Alphabet proposed recently, so my transcription is certainly at fault in places. For example, I believe the phoneticians would use a single symbol where I have used ng to represent the sound in English 'sing', and possibly capital R and small r for the sounds the Spaniards write with 'rr' and 'r' respectively. The apostrophe is for the hiatus that I believe philologists refer to as a glottal plosive.”

“Pray give me till this same time tomorrow to see where this may lead,” said Holmes.

“I shall,” said our visitor, and left.

Exercise: From looking at the first page of Mr. Bond's notes, which languages appear to you to be the most similar? Choose one possibility. We keep no record of your answers; you know better than we how to follow up

on your results. Makassar and Bugis Makassar and Saleyer Bugis and Saleyer Not enough evidence Think about how many of the forms on the sheet don't match up.

Comparison - A Dawning Light

When I came down to breakfast the next morning, Holmes was at his desk, a pile of foolscap in front of him and other sheets lying crumpled on the floor. “Well, Holmes,” I said. “Aren't you stirring rather early?”

“Rather late, Watson. I started looking at the lists of words Mr Bond left us. Though the languages show a great similarity, deuce take me if I can find any consistency to it! I've been at it all night.”

“Then I prescribe a couple of hours' quick sleep, followed by a hearty breakfast of eggs, ham, and coffee. I know you won't depart from the chase for any longer than that, but it will clear your brain. I'll ask Mrs Hudson to lay breakfast for you at half past nine.”

He made himself comfortable on the couch. When he slept, I made arrangements for his breakfast. Then I took up Bond's lists of words, and soon saw for myself that for whatever pattern I thought I discerned, an exception to it lay close at hand.

Holmes awoke in due time. While we breakfasted, he laid out his frustrations to me. “Beastly things, Watson, these foreign languages! Yet I suppose were our own English laid alongside Dutch and German, we should find the same tangle. There is a consistency here, I swear; but it's woven through so many threads that even the most careful note-taking fails to capture it. If only I could follow thirty or forty trails at once!

“For example, look at the first entry on the list. In all three languages the words begin with a p sound. From that we can deduce nothing about how the languages diverged, though it does suggest that the progenitor language used the p sound.

“Examples like ka'daro 'husk' and bunga 'flower', in which all the sounds match perfectly in all three languages, also tell us nothing about differences, but they do expand the list of sounds that occur at the beginning, middle, or end of words, especially if we find the same correspondences elsewhere in the list.

“Some of the entries don't seem to point to a common ancestor at all. 'Leaf,' for example, is leko' in Makassar, daung in Bugis, and taha in Saleyer. Since there are no similarities, it could indicate that the people Bond interviewed were simply thinking of different kinds of leaves. They might well have in their vocabularies words that match those of the other languages, but those words didn't come to mind in the situation. Or some of the words might be taken from still other languages, much as we English use words taken from French every day, to the disgust of the French.

“But the entries where one word is similar to the others, yet not identical to them, raise interesting questions of how the languages may have changed with time. Look at the entry for 'banana'; the sounds nt in Makassar unti match a lengthened tt in Bugis utti, where the n sound in that position may have taken on the properties of its neighbor. We don't know where Saleyer stands on that one, because they use a different word, but as we comb through more entries, that may become clear.

“The entry for 'grass' suggests the same sort of thing: Rea in Makassar and Saleyer, area in Bugis. Either Bugis has added a prefix a-, or an initial a has disappeared, so we must keep on the lookout for other evidence of the same process.”

“But Holmes,” I exclaimed, “you've said several times already that we need to track similar kinds of change through many cases, and every one of them may seem equally tenuous!”

“Tenuous, yes, my friend. But though one correspondence by itself be light as gossamer, enough similar cases may weave a strong fabric indeed,” said Holmes.

I answered, “Then that must be what kept you up all night, the thirty or forty trails you spoke of trying to follow.”

“Exactly, Watson. And now that you mention it, a new thought comes to mind!

“What if we could harness the power of Babbage's analytical engine to keep track of all that for us! Quick, write off some notes for yourself on what we

have been saying, while I prepare to make off with the whole manuscript to the Science Museum!

“You, meanwhile, will await our Foreign Office friend and explain to him that I have a lead that may indeed provide him a trustworthy answer. But I will have to strain my intellect to the utmost, if I am to bridge the gaps in our generation's understanding of the Analytical Engine, that it may become the means by which we reach our solution.

“The name of this person who is constructing an Analytical Engine from Mr Babbage's notes? I shall enlist his aid.”

“Hamilton Goodge is the name,” I told him. “A young man, well prepared in mathematics, and quite at home in his maze of levers and wheels. You know the type.”

So I remained in the upper-story flat we shared at 221B Baker Street, and was able to give Mr Bond the news that Holmes's chase was afoot, this time in the workshop of a museum rather than in the moors of Scotland or the opium dens along the Thames. I went over my notes with him on the specifics Holmes had already noticed.

And Holmes? Never a word from him the rest of the week. Bond went down on the Friday to see what was happening at the Museum, but Holmes refused to see him, so intent was he on formulating the exact instructions he must give the Analytical Engine in order to get the answers he needed.
Exercise: Which forms might tell us how the languages have changed? Choose at least one: husk flower leaf banana grass Some don't give evidence for anything. Some give evidence for what the ancestral language was like. Some give evidence for changes that have taken place.

Comparison - A Retrospection

Sherlock Holmes must have slept at the Science Museum if at all, because he did not again make an appearance until Monday.

“Look, Watson! We've got it!” he exclaimed as he threw his coat on the wicker chair. “Your friend Goodge is a marvel! He became as engrossed as I was in the problem. Furthermore he succeeded in communicating with Babbage's engine by punching holes in a certain way in pasteboard cards like those that control the Jacquard loom. The holes encoded instructions to the engine as to what to do.

“See this? From the lists of words, also put into the engine by punching holes in pasteboard, he performed a simple manipulation. For each of the entries, he stacked up the words in the three languages, then sliced them the other way to get the sets of correspondences, each with a note on where in the word it came from.”

Languages	Words		Correspondences	Position
Makassar	Rea		/ a /	first in word
Bugis	area	---	R r R	between vowels
Saleyer	Rea		e e e	next to last
			a a a	last in word

“Then for each correspondence in turn, we added a symbol to show tentatively what sound in the earlier language the three corresponding sounds might have developed from. We developed a shorthand for position:

for the first and last sounds in a word we used #_ and _#, the _ indicating where the sound in question stood vis-à-vis its neighbors. We also used C to represent any consonant and V any vowel, so that the middle two positions are recorded as V_V and _C# respectively.

“Some sounds seem to have come into the languages later in time, or to have dropped out -- we can't yet tell which way it went. We put a slash “/” to match where this occurred. Similarly, we put a full stop “.” in correspondences where a comparable form was not present in one of the languages.

“We noticed that some correspondences are quite common, like the a-a-a correspondence in the example. Others are quite rare. My friend Goodge came up with a mathematical formula by which a group of matching words could be recognized as highly consistent if every correspondence in the same group recurs plentifully in other entries, while if some of the correspondences are uncommon, his formula indicates a lesser degree of certainty.

“So we began with the strongest correspondences, and used the picture they gave of the earlier language as a framework within which to locate the less well attested correspondences. By this means, most of the correspondences turned out to fit the picture well.

“On the other hand, some correspondences fit not at all. We ended up regarding them as evidence, not for the historical development of the three languages, akin to genetic development, but nonetheless evidence for occasional dealings with speakers of still other languages like Malay and Portuguese, some of whose words passed over into the languages in the list in a most helter-skelter way.

“Based on the evidence of the most broadly attested correspondences, we can inform Mr Bond that without a doubt, Saleyer and Makassar diverged from Bugis before they separated from each other. Words with strong attestation like batu 'stone', bunga 'flower', and tunu 'roast' are identical in all three languages, descended unchanged from a stage before the two-way split.

“But by far, the evidence for the split is more interesting. There was an earlier sound that has come down in a number of words as “k” in Makassar and Saleyer, but “'”, the glottal plosive, in Bugis. It is found in forms like kutu and 'utu, both meaning 'louse', whose u-u-u and t-t-t correspondences are widely attested, so their plausibility is high.

“There is also a sound in Bugis that Mr Bond wrote as “i”, though I sincerely doubt that it sounds exactly like “i”. It invariably matches up with “a” in Makassar and Saleyer sampulo, Bugis sippulo, meaning 'ten'. It does not match with “i” as in lima 'five', which sounds the same in the three languages. I cannot explain it, but it is extremely regular.

“And what I suspected earlier, that an original nt in Makassar unti 'banana' may have shifted to tt in Bugis utti, is also confirmed by the words for 'ten', as well as by Makassar and Saleyer bintoeng 'star' against Bugis vittoeng.

“I needn't bore you with the details. The good Mr Babbage years ago developed a printing machine to go with his analytical engine, so the full details will go to Mr Bond. Please send him a telegram inviting him to stop by at his earliest convenience; then we shall wish him bon voyage.”

And thus ended Sherlock Holmes's foray into philology. Once he had tracked down his clues and turned them into deductions, he lost further interest in the Babbage Analytical Engine and returned to his real passion, crime. But I hope Mr Goodge continues from there, lest the fruit of his labours be lost to philology.

Note: Unfortunately for the science of language, neither Holmes nor Goodge carried on with their efforts. Nor was Mr Bond heard of again after he arrived in Singapore, and his superiors feared the worst. It was not until just over a century later that the linguist Joseph Grimes began to follow a similar train of thought, having the benefit of a marvelously improved descendant of the Analytical Engine. The results can be examined through wordcorr.org. For further information on the Babbage engine, Holmes, or Watson, consult Wikipedia.org. There is more detailed information on Wordcorr itself in the modules that follow. The actual language data are

from *Languages of South Sulawesi* by Charles E. Grimes and Barbara D. Grimes, from *Pacific Linguistics, Series D, No. 78*, 1987, with modifications in the representation of the words to fit the phonetic conventions of the late nineteenth century.

Behind the Scenes

Data into Wordcorr

Putting together a Holmesian story was fun. It didn't require much real work, because I already had a Wordcorr collection, JG-SulSel12, that included the three languages talked about.

To begin, I used the Connexions tutorial Installing Wordcorr to install Wordcorr on a new XP laptop, a Lenovo T43 ThinkPad. I had written other parts of these tutorials on one of the Toshiba laptops used to create Wordcorr itself; but after three and a half years of pretty intense work its hard disk was going bad.

At an early stage in the development of Wordcorr, I had already imported some WordSurv data into Wordcorr, the SulSel (Indonesian abbreviation for South Sulawesi) data on twelve languages of the region. Then I exported it to an XML file in order to test both operations. For fifteen years before that I had kept the same data under an old program, WordSurv 1, that had preserved the data for me but did not have capabilities for handling either the International Phonetic Alphabet symbols (IPA) or the comparative analysis.

So I already had a Wordcorr XML file containing the data and some earlier analyses. Before I imported it to the new computer, I had already verified the correctness of most of the IPA transcriptions on the older computer, and edited the faulty ones using Wordcorr's intuitive approach to typing IPA.

The Holmes View

Within the Wordcorr collection that I was doing serious work on, I put up a special view called "Holmes" that showed only the three languages I was focusing on and left the other nine out. (Wordcorr never destroys or loses data; but any new view can be set to bypass some varieties.) I set a threshold value of 60% for the Holmes view, to make sure that all correspondence sets contained information from at least two of the three languages.

Then I went through Wordcorr's basic Annotate-Tabulate-Refine cycle for the first hundred entries. You'll hear enough about that cycle later, so I'll resist getting you bogged down in the details here. It didn't take very long, a few entries one day, a couple of dozen the next, when other projects made me sleepy.

Finally I invoked the Summarize Evidence function from the Refine panel, giving it a cutoff of 0.0 on Frantz's measure of strength in order to filter out reconstructions whose component correspondence sets were only weakly attested.

Pinpointing the Relevant Data

At that point I settled down to scrutinize the patterns that were emerging. Starting with the best attested correspondence sets (those accompanied in the reconstructions only by other well attested correspondence sets), in about three hours I identified which forms were worth having Holmes call attention to.

Retrogression

Since the International Phonetic Association was just getting going around 1888, I retranscribed the raw data using the phonetic alphabet that Henry Sweet in London (the prototype for Shaw's Henry Higgins in *Pygmalion*, later *My Fair Lady*) and his colleagues in Paris were discussing at that time, with a little fudging on details they hadn't gotten around to yet. The IPA alphabet has been upgraded several times since then, but its essential character has changed little.

While I was doing the real linguistics, I also poked around in Wikipedia.org for all kinds of historical, political, computational, linguistic, and literary information, besides reading through a collection of A. Conan Doyle's Holmes stories paying special attention to their style. I found a suitable MacGuffin that got a sorry excuse for a plot going: the languages it was convenient for me to use are spoken just west of the Spice Islands, and British boiled mutton seems to have been even blander then than it is now.

And I hope you enjoyed reading it as much as I enjoyed writing it. Maybe you saw something there about language that you hadn't noticed before.

Wordcorr overview

Comparative linguistics looks at **vocabulary item**s from two or more **speech varieties** that might have a common origin and attempts to explain how their differences developed.

For example, the word for “cat” is pronounced gato in Spanish, gatu in Portuguese, gat in Catalán of eastern Spain, and gat in Occitan of southern France. (The forms given here represent the pronunciation, not the traditional spelling.) It's easy to see that the four forms might have a common origin.

It's also easy to guess that the ancestral word behind this **entry** may well have contained the sounds g-a-t. But to decide whether the original had a vowel at the end and Catalán (let's just call it C) and Occitan (O) lost it, or whether there was no vowel at the end and S and P developed one, we need much more data. And if S and P did inherit an earlier vowel in common, was it o or u? Or something else?

Word list entry		**and its**	**Correspondence sets**			
	'cat'		**S**	**P**	**C**	**O**
S	gato		g	g	g	g
P	gatu		a	a	a	a
C	gat		t	t	t	t
O	gat		o	u	/	/

Clearly it takes a lot more than a single set of words for “cat” to give evidence for the history of those speech varieties; but it is a start. When we get to comparing much beyond four varieties, it's easy to get lost, especially if you're making the comparison the traditional way using file slips. That's where the computer program called **Wordcorr** comes into the picture.

In a nutshell, Wordcorr lets you put together a **collection** of **data** from as many **speech varieties** (like Spanish and Portuguese) as you are investigating. You can start with just one **entry** like "cat," and build up to two or three hundred **vocabulary item**s and many more. Once the data are in, Wordcorr walks you through the many decisions you have to make about what the data actually show. As you work, it organizes and reorganizes the results of those judgments. That's how you put the whole picture together. With Wordcorr, it takes much less time than if you tried to do everything by hand.

--Here's a self test to assure you that you're on the right track. We don't keep any record of your answer, because you know perfectly well what to do with the results. Try it:
Exercise: A typical Wordcorr collection is likely to contain ... more speech varieties than entries That would imply very short word lists. more entries than speech varieties That's correct. about the same Typical collections would be much larger than that. We mentioned four or more of one, 200 to over 300 for the other. A collection consists of a manageable number of speech varieties, rarely more than a few dozen, and entries for all the data you have available.

Wordcorr tools

Here are some of the main things Wordcorr helps you with. There are other tools that will come up later, but these are the key functions.

Enter Data. Type in **word list**s for **comparative** analysis. The first word list establishes the sequence of **entries** for all the others. You have access to the complete **International Phonetic Alphabet** plus some other symbols. You can **edit** data after typing them in.

Define View. Since alternate analyses of the same data are possible, you perform each analysis in a separate **view**. You give each view a name, decide which **speech varieties** of the **collection** are to appear in the view and in what order, and the **threshold** percentage of the varieties in the view that must contain a **datum** in order to form a **correspondence set**.

Annotate. For each **entry**, you examine the data and decide which words could plausibly have descended from the same ancestral form (usually called the **protoform**, which you will endeavor to reconstruct). Each set of such words constitutes a **group**, and each group gets a one-letter **tag**. Within each group, you line up the **segment**s that match each other, perhaps stretching the shape of some words by typing in the **Indel** symbol "/" to indicate that either a new sound may have been inserted in some of the words that would otherwise match, or a sound may have dropped out at the point where you insert the symbol. There are some minor tools for annotation as well. All the words in a group must have the same number of symbols; but different groups in the same entry do not have to have the same number of symbols.

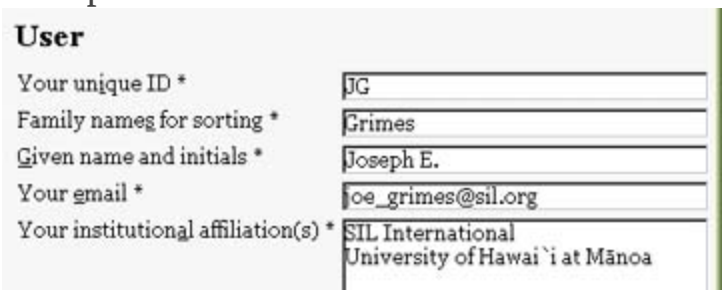
Tabulate. Once all the groups in an entry are annotated, the groups with enough speech varieties entered to meet or exceed the view's threshold are taken up one at a time. Wordcorr transforms each group into **correspondence set**s. For each correspondence set, you indicate the probable **protosegment** of the **protolanguage** that the set represents, and indicate the phonological **environment** in which it probably was situated within the protoform. At first these judgments are only educated guesses; but as you proceed, patterns you begin to recognize take shape to give an attested picture of how changes happened.

Refine. As soon as a group is tabulated, its correspondence sets appear in the view's **Results** structure, which is organized **phonetically**, by the **place and manner** of articulation of the protosegments. You can **refine** it there, moving things around, changing protosegments, environments, and orderings. You can also extract a summary of the evidence for your hypotheses about language change, print out the whole results structure, or extract just the raw data.

Exercise: The word list data that you type into Wordcorr end up as _____ in the _____. (Click on the two answers in the order in which Wordcorr provides them.) varieties Varieties are part of the data, not of the analysis. protolanguage As you do the analysis, you define the protolanguage. results Indel Indel is an annotation symbol that guides the analysis, so it's related to the correct answer. correspondence sets group tag Group tags guide the analysis, so they are related to the correct answer. The outcomes of annotation and tabulation are kept for you in permanent form. The results from which you will deduce the form of the prolanguage and the changes that have taken place in the speech varieties consist of correspondence sets organized by protosegment and environment.

Information about you

After you install Wordcorr, it needs a small amount of information about you. This is the same kind of public information that you would include with an article published in, say, *Language*. Wordcorr will attach this to your data collections automatically. Here is what Wordcorr asks for on its User panel:



User information from Wordcorr's New User panel.

1. A short personal identifier, 1 to 3 capital letters such as your initials (like JG). When we start registering people who use Wordcorr, by means of the Internet, wordcorr.org will accept only identifiers that aren't already registered to someone else; so at that time you might have to change.
2. Your long term email address
3. Your first and last names -- given name(s) and family name(s) separated, to facilitate alphabetizing
4. Your institutional affiliation. If you have more than one, put each one on a separate line.

You enter this obligatory information only once: when Wordcorr first comes up on the screen. From then on you have to worry only about making changes (like a new email address) when they come up. Click the File menu, choose New User, and fill in the blanks.

Once the information about you is filled in, you can start to create or import a Wordcorr collection.

The shape of collections

Everything Wordcorr does is in the framework of a **collection**. You may have more than one collection, one for each family of languages you are working on. If your focus is research, you may spend most of your time on a single collection. If your focus is teaching, you are likely to keep a number of collections around for your students to work on.

A collection consists of

- two or more possibly related speech varieties (like Buduma, Cibak, and so forth); you can't make comparisons with fewer than two
- one or more entries, each containing words of about the same meaning (like words for 'cat' in one entry, 'sleep' in another, and so forth) ideally with data from each of the varieties, and the more entries the better
- a full name for the collection (like "Chadic-Biu-Mandara")
- a short name for the collection (like "ChadBM")
- information about the collection as a whole
- information about each speech variety
- a file name under which the collection can be stored (usually the same as its short name)

You have two main interests in the collection:

- the raw data, written phonetically (or phonemically if you are confident about the phonologies)
- various analyses you make of the same data

By the time you are ready to interact with other linguists, you will develop some other interests:

- documentation for where the various pieces of data came from, published or unpublished
- accurate information on each of the speech varieties, either taken from the Ethnologue or suitable for inclusion in the Ethnologue
- names and Wordcorr IDs for every linguist who has worked with you
- names or other identifiers (like "70 year old man from Kabundu") for every speaker of one of the languages you cover who have given you

data

Collection metadata

Identifying a Wordcorr collection is a lot like identifying a book in a library. Some of the information, as might be expected, is more aligned with electronic libraries than with traditional book publishing.

Here are the different kinds of metadata (data about your data) that Wordcorr asks you for on the Collection panel:

- Names of the collection
- Creator and collaborators
- Glossing languages
- Description, key words, remarks
- Published sources
- Geographic coverage
- Permanence

Collection titles

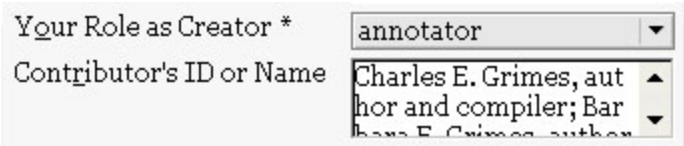
Title and short title on the Collection panel of JG-SulSel12.

Every collection has a **title** that is useful for other linguists, like Austronesian or Southern Utoaztecan. The title indicates the spread of languages you are investigating.

You prefix your Wordcorr ID to the title with a hyphen. This heads off a potential bad situation where thirty Wordcorr users have thirty different collections all called “Austronesian,” and some of them are trying to collaborate with others.

Wordcorr asks you for a **short title** for the collection, which also carries your Wordcorr ID prefix. This usually makes a good file name too, when you convert your Wordcorr database to an XML file for safekeeping or collaboration with another linguist. So a JG-Austronesian collection might have a short title like JG-AN that would readily identify it to any specialist in Austronesian languages.

Collection creator and collaborators



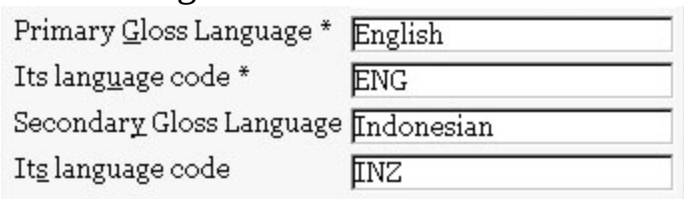
Creator role and collaborators from the Collection panel of JG-SulSil12.

In terms of internationally established metadata categories, you are the **creator** of any collection you initiate, and you are a **collaborator** on any collection that somebody else initiates and you contribute something to. Anybody else who contributes to your collection is a collaborator too. For example, if you make additions or corrections to the data in somebody else's collection, or if you set up a separate view to do your own analysis, or modify an existing view, you are a collaborator.

Being a creator doesn't necessarily mean you did all the field work or the analysis or paid the bills. There is a standard list of **roles** from which you can choose the role for the creator of a collection. I am the creator of JG-SulSel12 in the example, but my role as creator is "annotator" because I took data from a published source and reworked it to fit the form Wordcorr requires, though I did none of the field work. Yet if I hadn't planned it all out and set it up, the Wordcorr collection wouldn't exist. This particular work is my baby. I'm not an author of the work I took it from, though I got permission from the authors ahead of time to use their data in this way. Furthermore, in the collection's metadata I give full information on the published source, so there is never any question about who did what.

Note: The Dublin Core is the basic international schema for metadata on published works. OLAC (the Open Language Archives Consortium is an extension to the Dublin Core specifically for linguistics.

Collection glosses



Two glossing languages, from the Collection panel of JG-SulSel12.

In order to work with many languages at once, we need to choose one or two languages in which we are likely to publish our results. The **glossing** languages are used to identify the approximate meanings of words in the languages whose relationships we are studying.

Comparative phonology lines up words of similar meaning and similar form. A Wordcorr entry might consist of the words for "house" in twenty languages that might be related, for example, and another entry have in it the words for "horse" in those same languages. And it might turn out that in one of the languages the "house" word doesn't match, but the "shack" word matches the "house" words in the other languages.

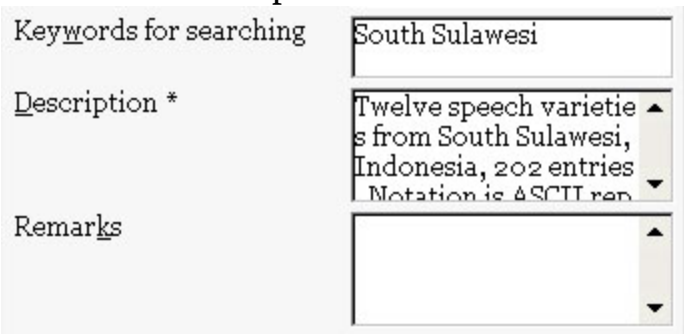
Word lists are not dictionaries. They are imprecise about exactly what the words in them mean. And because the meanings of words change over time just as their forms do, what we need for comparison had better not exclude words that could have come from the same ancestral form as words in other languages that now mean something a little different.

With each glossing language name goes its Ethnologue code.

Note: The codes that identify languages, both the languages of description (the glossing languages) and the languages being compared (the subject languages) have recently had some codes modified by international agreement. Verify codes against the Fifteenth Edition of the Ethnologue.

As a quick clue, codes in capital letters, like the ones in Figure 1, are from the Fourteenth Edition or older. Currently valid codes are in lower case.

Collection description



Key words, description, and remarks from the Collection page of JG-SulSel12.

Names and titles are often not informative enough that a person who might profit from the contents of a collection can get a clear idea what is (and is not) in it. So Wordcorr requires a paragraph or more to **describe** the contents. It doesn't have to be long, just long enough to be useful.

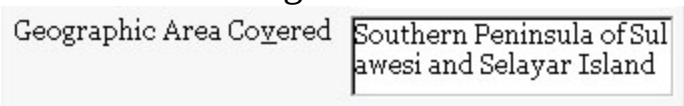
If you were doing a Google search for your collection, using just a regular Web browser like Firefox and not a special search engine developed for linguistics, what words would you use in the search? Put those in the Keywords field. Think about what other **key word**s might lead people to your collection too. Separate the key words (or phrases) with comma space.

There is also an optional **Remarks** field for essential information that doesn't fit standard library categories. For example, early versions of WordSurv, and many collections imported from spreadsheets, were made at a time when linguists had no access to character sets for the International Phonetic Alphabet (**IPA**) in Unicode. So linguists each made up their own private conventions, things like capital "N" for angma, capital "E" for epsilon, "q" or "7" for glottal stop. There was no way they could standardize among themselves. So any collection with that kind of history needs an explanation of the original symbols in its Remarks field. From experience, we recommend that you put it in even if you think you've gone

through and converted everything to IPA – it’s not impossible that you might have missed a few.

Anything else that linguists who use your data fifty years from now should know about your collection also goes into Remarks.

Collection coverage



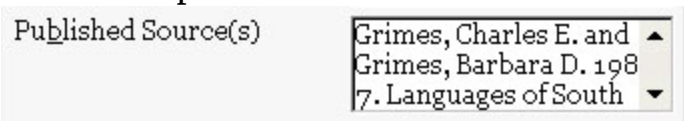
Geographic distribution of languages included in JG-SulSel12, from the Collection panel.

Geography is a factor in comparative linguistics. Some languages are spread all over the world. Others are used only in one valley on a small island. Language boundaries may or may not coincide with political or geographic boundaries at any level.

Back in the days when comparing even a couple of dozen languages was a lifetime task, because everything had to be done by hand, there was a temptation to classify the intermediate levels of the language family tree by geography. So one sees references to languages of Borneo, or Formosan languages, or languages of the Grassfields, in the comparative linguistic literature. And it is true that there are places where linguistically close varieties live next to each other and linguistically distant varieties don’t.

But finding that a geographic region contains a single family of languages ought to be a conclusion of comparative research, not a premise. The geographic aggregation of languages may or may not match linguistic history; it is always suspect until we’ve done the actual comparison. But we still want to know where the languages are spoken.

Collection published sources



Published source from the Collection panel of JG-SulSel12.

The **word list**s that make up a collection come from a variety of sources:

- Published word lists or dictionaries
- Word lists you and your associates have collected
- Word lists other linguists have passed on to you
- Notes on languages made by nonlinguists
- Documents in some of the languages

Language surveys may cover areas where little is known about what languages are spoken there (yes, there are still parts of the world where this is true). Even in such situations, it is profitable to include word lists from well documented languages of the region as well as the lists the survey picks up. The better documented languages might or might not be related to some of the languages included in the survey. You want to know that, in order to tie the survey data in with comparative work already done.

Published sources cover, of course, anything modern libraries can catalogue with ease that contains data you use or discussions of comparative results:

- Books
- Articles
- Microform documents
- Archived paper documents
- Electronic documents (some, not all)

Looked at another way, anything you can write a bibliographic entry for that would be acceptable in a refereed linguistics journal can be considered

a published source.

What about Web pages or computer files as published sources? That's still a problem. There isn't enough experience among linguists for a consensus to form. Web pages come and go all the time. Nobody likes to be confronted with the dread Error 404 "Page cannot be found" message when all the World Wide Web's technology draws a blank on getting a page you asked for.

Permanence is the issue when deciding what can be cited. The Web page you are looking at in this tutorial has been designed to be accessible fifty or more years from now. You can see several acceptable ways to cite it by clicking on "Cite this content" at the bottom of this page. A good rule of thumb is

> **Note:** A Web page to which a well funded university, museum, archive, or library is committed for the long haul is citable. Anything that is posted on a personal Web page, even that of a reputable linguist, is probably not citable.

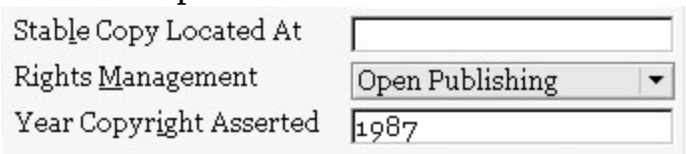
What you want is to archive your material on a Web site that is committed to permanence. Make arrangements with one of the linguistic archives, on the order of AILLA, DOBES, or PARADISEC. Or find a university or museum library committed to the preservation of electronic data. Some national archives may be set up as repositories for scientific data.

Keep in mind that while "endangered languages" are very much on people's minds, far fewer people are conscious of the problem of "endangered data."

Another question that comes up is whether a published source should be listed as metadata for the collection as a whole or as metadata for one of the speech varieties. The answer depends upon its scope. In one collection, two comparative works are given full references under the collection, because neither is restricted to a single variety in the collection. Then under the

specific speech varieties, in the published source field, all that is needed is the author-year type of citation: *Savage 1986*, *Reid 1971*.

Collection permanence



Information from the Collections panel about permanence of the collection.

The question of permanence has already come up in regard to what can be considered a published source in the age of digital archiving. Digital data repositories in universities, museums, and national archives are likely to be able to furnish your data to scholars after more than a century. Personal Web sites and your computer's hard disk are not.

Once you have a permanent home for your collection, you can enter the name or URL of the archive in the collection itself as "Stable copy." In the example above, the location of the stable copy is still up in the air. Eventually it will need to be settled.

Who has the right to allow reproduction of a work? That's a broader question than who holds the copyright.

If you share your Wordcorr collection with a colleague, and the colleague fills out a new analytical view of the data in the collection, the collection with the new view's analysis added is a derivative work — a hefty intellectual project that takes your original collection as its point of departure. It builds on it in the tradition of good scholarship. But it's still your collection, modified by someone else with your knowledge, and quite possibly with a good deal of communication between the two of you.

New questions of ownership have developed as computer usage grows. Computers create, store, copy, transform, and transmit information electronically. The medium opens up a potential for scientific and literary

collaboration that did not exist when paper was about the only vehicle conceivable for writing, storing, copying, or disseminating information.

So it's not surprising that it's the computer people who have put a lot of effort into clarifying ways in which the rights of each person involved in creating, modifying or using information can be kept clear. Their efforts fall under the general rubric of rights management.

One of the best ideas that has come out of considerable discussion is that copyright is more than a means of preventing people from reproducing your work. It’s perfectly proper for you as copyright holder to permit, and even encourage, other people to develop what you own further.

This has developed into the idea of **licensing** the use of copyrighted material. And more and more nowadays, under the impetus of the Open Source movement, the license is granted without charge – provided you meet certain conditions that ensure that the information about you as originator of the document, and about changes in it that others have made, is sent along as part of the document itself. This tutorial, as you can see for yourself, is released under an open Creative Commons license.

Wordcorr gives you the option of choosing among three types of open license: Open Publishing, Common Content, and GNU Documentation.

You don’t have to copyright your collection. But if you do decide to, you want to put the year in which you assert copyright into the collection metadata. Then you can register it when it’s convenient.

Note: Things change constantly in rights management. Consider this tutorial probably out of date, and check around the Internet and in your institution to see what possibilities fit your situation best.

The shape of speech varieties

Languages and Dialects

Wordcorr helps you compare different speech varieties. People often want to know if these are "languages" or "dialects." But it makes better sense to think about that distinction after we have done the analysis, rather than making it the starting point.

Note: Many people think "languages" must be the way powerful people speak, and "dialects" must be the way people without power speak. From the linguistic side, however, a more useful distinction is whether people who speak differently can or can't understand each other inherently, without going to the trouble of learning each other's languages. Looked at linguistically rather than politically, if people who speak differently understand each other solely on the basis of the similarity of their mother tongues, they speak different dialects of the same language, whether the dialects are prestigious or not. And if they don't understand each other, they speak different languages. The implications of this distinction for education are far-reaching.

So when the ordinary speech of two social groups is different enough to notice, we can say they speak different **speech varieties**. That's a usefully neutral term -- we can use it without having to argue about whether the varieties should be called separate languages or dialects. Later, we should have enough information to state the ways in which they differ, and how each got to be the way it is. At that point both politics and inherent intelligibility can reasonably come into the picture.

Wordcorr comes in at the stage of identifying the systematic, regular differences among speech varieties.

What We Track

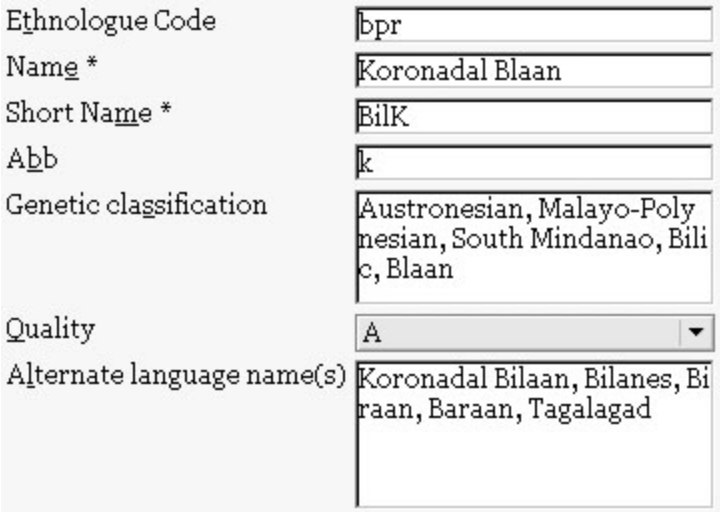
Some of the things that need to be kept track of for each speech variety are

- There is a standard set of codes for nearly all the world's languages, reducing confusion when one speech variety has different names in different countries
- Different languages may have the same name
- Many languages have more than one name
- Languages may also have conventionalized short names (somewhat like nicknames) and abbreviations
- Languages are often identified by the place where they are spoken
- Word lists in a collection may come from both published and unpublished sources
- The transcription of every word list needs to be explained unless it's true IPA

Wordcorr collections are arrangements of parallel word lists from speech varieties you want to compare. Accordingly, you have to give Wordcorr enough information via the Varieties panel to not get any two of the varieties mixed up. This information that is kept in Wordcorr, or **metadata** (data about data), is also the key to letting other linguists know what is in your collection, so that you can possibly get together and collaborate on the analysis.

Note: We can thank William O. Bright, editor of the first edition of Oxford University Press's *International Encyclopedia of Linguistics*, 1992, for insisting on the term "speech variety" as replacement for the traditional "language" vs. "dialect," which often generated heat rather than light.

Variety names and codes



Names, abbreviations, and language codes from the Varieties panel.

Variety Names and Short Identifiers

It's hard to identify languages precisely, but for comparative work, we don't dare get them mixed up. Wordcorr goes out of its way to help you get them identified unambiguously.

Every language has a name. It isn't necessarily what the speakers of the language call it, because the speakers of some languages use phrases like "how we talk" to refer to their language, some names are considered unpronounceable by most outsiders, some are names that outsiders use, some are invented by linguists and the speakers don't recognize them, and some names (cited in quotation marks in the Ethnologue) are uncomplimentary or even obscene names bestowed by the neighbors. Many languages have more than one name.

The obligatory and unique Name of each language is preferably the name the speakers use for it (if they name it and it isn't too awkward for the rest of the world), and otherwise, the main name that the Ethnologue gives. For speech varieties that aren't in the Ethnologue, try the name of a related

language followed by "of [place-name]" (like "Zapotec of the Isthmus") to distinguish it from other languages in the vicinity.

Note: The *Ethnologue* is the primary source for information about language names and codes worldwide. There is an index of language names that takes you to the Ethnologue's main name and code for all the languages it lists. If you don't know the main name, that page tells you how to search further for it.

Linguists who study the languages of an area usually develop their own shorthand for language names. Find out what that shorthand is for each variety in the collection and use it as the obligatory and unique Short Name. If you're working with speech varieties that are different from any that other scholars have written about, make up your own. Short Names are ideally four characters or less, because they are used in Wordcorr's Summary of Evidence and longer Short Names could take up too much room.

Note: Collections cannot have two varieties with the same Name, or with the same Short Name.

Language abbreviations (abbreviated "Abb" on the Varieties panel) are one or two characters that identify the variety. They are a holdover from an older program called WordSurv. It used a single character (treating upper and lower case as different and admitting numbers and punctuation symbols to give space for 80 varieties) as the main identifier for each variety. If you use a converter to import a WordSurv collection, the abbreviation will appear under Abb. Otherwise abbreviations are optional.

Many, perhaps most, languages have more than one name. The main name goes into Name. All the others, including uncomplementary ones in

quotation marks, go into Alternate Language Name(s). You can cut and paste from the language entry in the Ethnologue.

Language Codes

Every known language has a language code. Most are based on the language name, to give the codes some mnemonic value. The codes were invented and assigned by computer for the Ethnologue, to give a means of relating the same language across country boundaries, whether its name is the same or different in the different countries.

Recently there has been a move to consolidate the Ethnologue codes with other codes. There is an international ISO 639 standard that gives two- and three-letter codes for several hundred languages. In addition, Linguist List has developed codes for languages long extinct like Elamite, constructed languages like Klingon (the Ethnologue already has entries for languages like Esperanto that are constructed but have acquired native speakers). So the new version, ISO 639-3, is basically the Ethnologue codes, expanded, and with some adjustments.

Suppose one of your speech varieties isn't in the Ethnologue? For starters, give it the same Ethnologue code as a similar language. There's nothing that keeps a code from being used for more than one variety. If you find out that none of the other varieties understand it, then get in touch with the Editor of the Ethnologue and lay out the evidence for getting a separate code for it.

Other Things About the Variety

Genetic classification is included here because sometimes it acts as a differentiator for languages that have the same name. You can get it from the Ethnologue language entry.

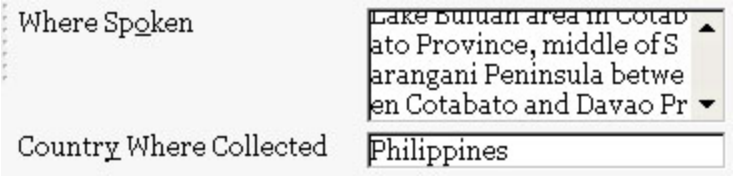
Of course, figuring out what the classification really is may be your major goal. That's especially true if you're involved in language surveys or if you are working in an area where the comparative work is conjectural, incomplete, or unsatisfactory (and there are some). In that case, include

only the upper branches that you're fairly sure about, or just leave the field blank.

If Wordcorr gives you convincing evidence that an existing classification needs to be changed, talk it through by email with the Editor of the Ethnologue.

The remaining field, Quality, invites you to express a judgment about the quality of the word list for this variety. Following the U.S. conventions for grading, A stands for a word list of the quality an expert linguist with experience in related languages can be expected to give, as good as you can get. B is competent but not stellar. C probably contains a few errors, but not enough to render it useless; you'll probably have to write in a lot of Remarks using the Edit Datum button on the Data panel. F means it's hopelessly bad, but you're including it anyway just in case something sheds light on something. (Edward Sapir had such a word list from Huichol in his 'Southern Paiute and Nahuatl', but he made a good point out of it because there was a regularity to the mistakes that allowed him to make a correct guess as to what some of the forms probably were.)

Variety locations



Geographic and social location from the Wordcorr Varieties panel.

Three spaces are important for locating a speech variety:

- Geographic region or regions
- Social groups
- Where the data were collected.

Geographic

Most speech varieties are associated with a geographic region; you can draw a line around where they are spoken on a map. Often the area of one speech variety overlaps with that of another.

Some speech varieties are not so easy to pinpoint in that way. Nomads, for example, move seasonally within a wide area, like Bedouins in the Middle East and Fulani groups in West Africa. Dispersed enclaves such as Mennonite Low German are dotted all over the former Soviet Union and the Americas. But for people to use a language actively, they have to be in some kind of community; they live near others who speak as they do.

Most of the world's speech varieties do occupy an area with more or less recognizable boundaries. But the boundaries may or may not coincide with those of a natural geographic region such as an island or a part of a river valley. They may or may not coincide with any political boundary; the colonization of the world has left many speech varieties divided by boundaries imposed only in the last five hundred years, because the speakers were there long before the colonizers came with their maps.

Some speech communities have moved around en masse. Others have split up. Others have stayed put. Others have merged into other groups and abandoned their original languages, like the Welsh in America.

And if one asks where English, Japanese, Spanish, Portuguese, some varieties of Chinese, Arabic, French, or Swahili are spoken, the answer is not just a long list of places. It is a moving target because the picture changes constantly.

What you want for the "Where spoken" information for Wordcorr is as succinct a statement as you can make. For very small speech varieties you might want to give latitude and longitude. For others you might place them with reference to a well-known city or a river, or within a province or a country.

And for the ones that are spoken all over the place, just do your best. It's okay to just cut and paste what the Ethnologue says if you don't have better information.

Social

In some places different groups of people in the same community use different mother tongues. Think of the varieties they use as connected to social groups rather than to geographically defined groups.

In most metropolitan cities these days, many languages are spoken. There may be sections of the city where different languages dominate - small enclaves. Or the members of a social group may be scattered around the city, but come together on special occasions.

The first place I lived in Honolulu, on our block you would hear mainly English and Hawai`i Pidgin. But one apartment used Samoan, and those who lived in the Buddhist temple on the corner used mainly Japanese.

In other places, the use of speech varieties may be highly institutionalized. Some caste groups in India speak differently than their neighbors with whom they deal every day, but the neighbors belong to other castes. In parts of the Amazon River basin there are communities where marriages are

always between people of different mother tongues. In situations like these there is often a high level of multilingualism.

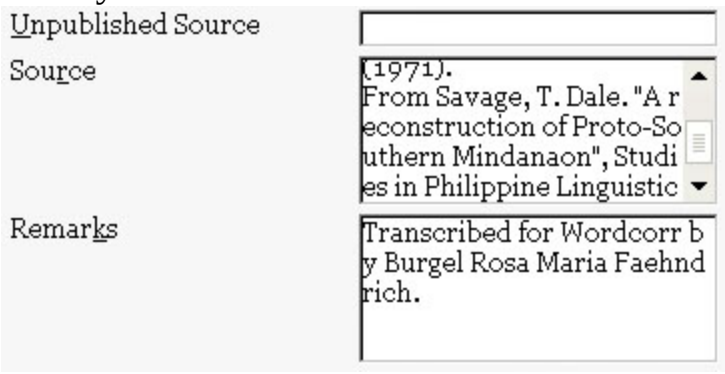
So statements about where speech varieties are used may require placement in social space as well as in geographic space.

Where collected

Linguists don't always have access to the actual communities that speak the languages they want to study. The only accessible speakers may be at a university far from their home, or in a refugee camp, or working as taxi drivers or bowling alley employees or janitors in an urban setting, or a patient in a hospital, or a canoe crew resting for the night at a stopoff point on a jungle river, or ...

That's why we ask about where the data were actually collected. Language data picked up from one individual in an airport departure lounge may require a trip to the other side of the world to verify or expand. "Country where collected" is worth knowing about.

Variety sources



Sources of information on one variety in the Varieties panel of Wordcorr.

Comparative word list data come from a variety of sources. Many linguists do their own field work, and get their word lists direct from native speakers of the varieties. Or information on different varieties may come from different books, monographs, or articles.

Or both. Before you go to the field you may search the library for words from languages in the area you are going to visit, and set them up in Wordcorr. Then when you get there, you may expand or correct that information with unpublished information that you pick up on the spot, and you may add in varieties that have not been reported on by any other linguist.

Other linguists are going to want to know just where the information on each variety came from. So you fill in the information as you start to work on each variety. If you need to augment it as you get more information, you can do that.

Unpublished sources

Some of your data do not exist anywhere else in computer accessible form or even in print. Apply the journalist's "what, who, when, where, why, how" criteria judiciously to fill in this information. If one of your students were to

visit the area two years from now, could they find the place where you collected your data? Could they find the person?

Remember that the people who give you information may not be in a place where streets or street addresses have reached. I've collected data sitting in a canoe in the middle of a jungle river in Suriname. But I could find my consultant (if he's still alive) tomorrow, because I asked for and got quite a bit about his position in his society.

Names can be a problem:

- Names may be closely guarded secrets
- Names used openly in tense political situations may lead to reprisals
- Names may change as people go through different phases of life

What is important for the study of language change is not so much the names and identities of individuals themselves, but the sectors of society whose speech those individuals represent. Record

- Age (often estimated or pegged to some datable, memorable event)
- Generation (child, unmarried young adult, childrearing generation, grandparental generation, ...
- Sex
- Office held in local society when interviewed
- Clan, lineage, caste, or other social segment that might differentiate speech
- History of contact with other languages

Published sources

Much of what needs to be said about published sources for speech varieties has already been given in the discussion of published sources for collections. Full bibliographic identification includes

- Author(s) or Editor(s)
- Title
- Series (if part of a series)
- Place of publication

- Publisher
- Year of publication
- Location of information within the work
- Location of information within an archive, if relevant

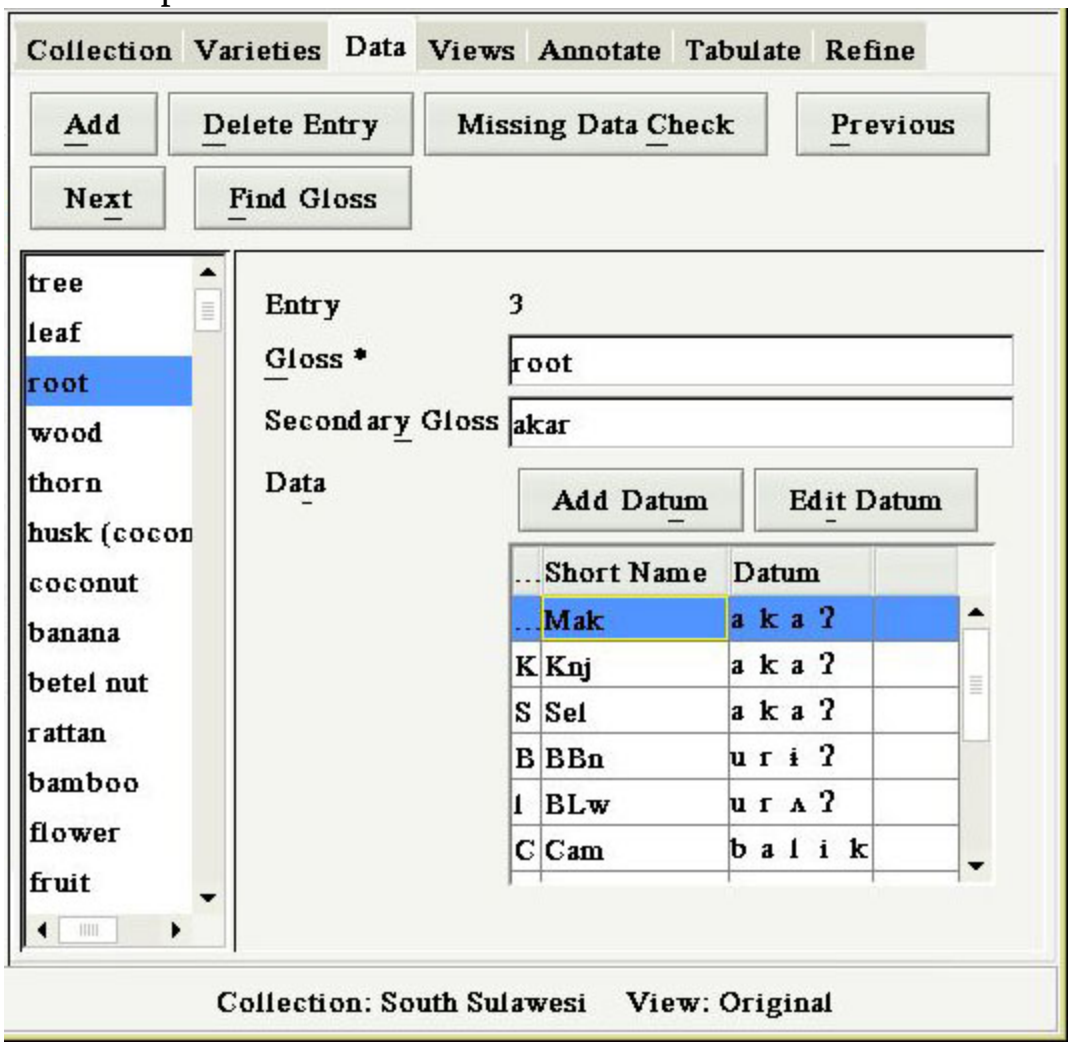
Remarks on sources

There is almost always some information that doesn't fit any of the standard bibliographic categories, or the electronic metadata specification, but that needs to be recorded somewhere. That's what the Remarks section is for.

Field notes, for example, probably use nonstandard, idiosyncratic notations other than IPA, especially if they were transferred to a computer before about 2005. The Remarks field for each speech variety is the place to explain the notation; Remarks throughout Wordcorr are all activated to allow typing in complete IPA using Wordcorr's easy-to-use conventions laid out in the Help facility.

Other things can go in Remarks as well: information about how to get to where the language is spoken, population estimates, information about where the data were collected if it wasn't in the homeland of the speech variety, common misunderstandings outsiders have about the variety or its speakers -- anything that might be of help to a colleague who might follow up on your work in the next few decades.

The Data panel



The Wordcorr Data panel

The Data panel is one of the seven action panels you can select by clicking on the tabs at the top of the Wordcorr display. It is where you type in and edit the word lists that contain the raw language data you plan to compare.

The upper part of the Data panel has to do with managing **entries**, which contain words that mean approximately the same thing (like words for "canoe") in all the speech varieties in the collection. The lower part has to do with managing the individual data items for each variety.

Note: For linguists, "data" is a plural count noun, as in "the data are correct," with "datum" as its singular, as in "one datum is missing." For computer people, "data" is a singular mass noun, as in "the data is correct," with "datum" as the corresponding singular count noun and "datums" as the plural count noun for several pieces of data. Moral: it pays to be bilingual.

So the buttons in the upper rows let you work with entries:

- Add: start a new entry at the end of the entry list, with obligatory primary gloss and optional secondary gloss
- Delete Entry: delete the entry currently selected; in Figure 1, it's entry number 3 for "root"
- Missing Data Check: tell whether any varieties, including the ones you can't see in the display in the lower right, are still not represented by data
- Previous: make the previous entry in the entry list on the left the current entry
- Next: make the following entry in the entry list on the left the current entry
- Find Gloss: find an entry by locating its primary gloss if you can't see it on the entry list

The buttons in the middle right of the panel let you add and edit a datum for a specific variety:

- Add Datum: select a variety and type in a datum for it
- Edit Datum: select a variety and edit the datum already there

Wordcorr entries

Comparative linguists group the data they work on into rough meaning categories. Each category is represented by a single word or phrase in the primary glossing language.

That doesn't mean that the words listed under a single gloss mean exactly the same thing in one speech variety that they mean in another. The fact of the matter is that words and phrases change over time in both form and meaning, and we want to know about both kinds of change. But when we're making our initial roundup of words, we have no way of knowing where they really belong. And because we can move them around later, we don't lose sleep over exact matchups of meaning.

Note: Early on in the development of the Austronesian language family, the word lima probably meant "hand" or "palm of the hand." But as languages diverged, the meaning shifted around in some languages to "five," "sole of the foot," "level surface" -- all perfectly plausible meaning shifts. If one were starting the study of Austronesian from scratch (not a good idea), the descendants of lima would pop up in several different entries because of the changes in meaning. Eventually, however, you would probably notice the similarities in form and the relatedness of the meanings, and pull them together into a single entry.

You start a new entry by clicking on the "Data" tab near the top of the Wordcorr screen. It brings up the Data panel. At the top of that panel you click the "Add" button. A blank entry comes up with a number one greater than the last entry you put in.

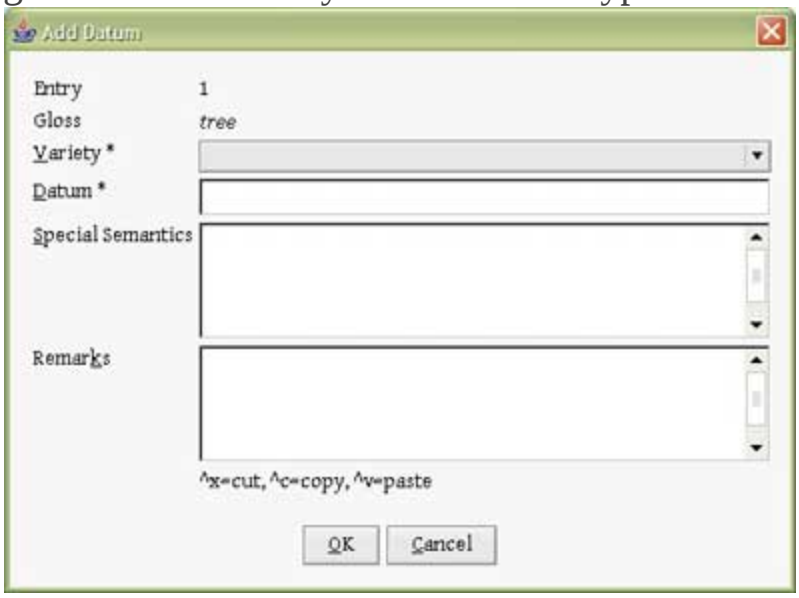
You fill in a primary gloss -- a word or phrase that suggests the meaning of the data you will be entering -- in the primary glossing language of the collection. If you have specified a secondary glossing language, you put in an equivalent word in that language too.

In the Data panel shown in the preceding module, Entry number 3 had a primary gloss (English for the South Sulawesi collection) of "root" and a secondary gloss (Bahasa Indonesia for the South Sulawesi collection) of "akar." That sets up the entry to take data.

Data from the keyboard

The keyboard is the most direct way to input data to Wordcorr. You can use the characters you usually use on your computer. You also have full access to the International Phonetic Alphabet (IPA) through your keyboard. And you can import data from other Wordcorr users, spreadsheets, and some other computer programs. This module deals with ordinary typing.

Once you have an entry started on the Data panel, with a primary gloss (and a secondary gloss if this collection has a secondary gloss language designated), you shift your attention to the middle of your screen. There you see the "Add Datum" button, and below it, what starts out as a blank grid where the data you are about to type will reside. Click **Add Datum**.



The Add Datum dialog box, from Wordcorr's Data panel.

The entry number and the primary gloss are carried in from the entry you were on when you clicked Add Datum. Below that, you fill in the obligatory **Variety** field by clicking on the darkened bar, which shows you a list of the varieties in your collection. Click on the variety whose datum you are ready to type in.

Next, click inside the **Datum** window (also obligatory), and type in one form for that variety in that entry. We'll pick up later on Easy IPA, Wordcorr's simplified strategy that gives you quick access to every current IPA symbol.

You can have more than one word in the same speech variety that matches the meaning of the same entry. To begin with, you may not know which one of several possible words reflects the forms you have for the other varieties. Wordcorr doesn't mind if you put in several forms for the same variety; some of them may have a common origin with one group of varieties, and another may match a different group of varieties. But activate Add Datum separately for each form.

There are two more text areas on the Add Datum dialog box. **Special Semantics** is what you use when the meaning of the word in this variety differs noticeably from the meaning implied by the gloss for the entry as a whole, you put a note in Special Semantics. The French word that matches the word for "house" in other Romance languages is chez, a preposition, not maison, the regular French word for "house." So you would put something like "where X lives, preposition" in Special Semantics.

Remarks is an important field. It holds all the information, or conjecture or guess, that Wordcorr doesn't give you a box for, but you feel it may turn out to be important anyway. Write all you like. Nearly everything in Wordcorr has its own Remarks field associated with it. This Remarks field applies only to the single datum you typed in the Datum field.

Then, when you want to keep what you've got, click the OK button at the bottom. Later, if you want, you can come back and edit what's there any time, and fill in Special Semantics or Remarks if some new idea comes up.

And if you decide you're not ready to commit what you've typed at least tentatively, click the Cancel button. You'll never see what's in that particular Add Datum dialog box again. If you come back to it, start over from scratch at the Add Datum button on the Data panel.

Data editing

Oops! Made a mistake! How do I fix it on the fly?

Your first line of defense, as with most computer programs, is your mouse or the arrow keys, together with the Backspace and Delete keys on your keyboard, which take out the character before or after the cursor respectively.

If you select a block of characters to work with, then you also have the old standbys Control-X for cutting out the selection and putting it on the computer's clipboard, Control-C for copying the selection to the computer's clipboard without cutting it out, and Control-V for pasting in the latest contents of the clipboard where the cursor is, or where the selection is. There's a reminder near the bottom of the dialog box: **^x=cut, ^c=copy, ^v=paste**

There's also an **Edit Datum** button on the Data panel. First you have to select the datum you want to change by clicking on it in the grid below the button. Then click the button. The Edit Datum dialog box that comes up is the helpful twin of the Add Datum dialog box. It's filled in with whatever you've got already. After you've edited, click OK. If you click Cancel, Wordcorr will retain whatever it had before you opened the Edit Datum dialog box.

Wordcorr isn't a word processor; it has more specialized things to do. So there's a tradeoff. You can type IPA faster in Wordcorr, and probably more accurately, than you can with your favorite word processor. You can't drag and drop to edit. But you can edit.

Data with phonetic characters using Easy IPA

You can type anything the International Phonetic Alphabet offers without having to lift your hands from the keyboard. Here's how Easy IPA works:

1. Type the character from a to z that looks most like the character you want.
2. Type Control-space by putting your little finger on the Control key, and while holding it down, tap the space bar once with your index finger.
3. Look at the popup box that contains all the IPA characters that are kind of like the one you typed in Step 1.
4. Choose the IPA character you want.
5. Type the number or letter associated with it.
6. The IPA character replaces the a to z character you typed.

That's Easy IPA in a nutshell. Say you type an "a" to get an a-like vowel symbol such as ash (also called digraph). Then you type Control-space. Up come all the a-like characters:

Characters kind of like a in Wordcorr's Easy IPA table.

Ash is indexed as "2", so you type 2. It replaces the "a" you typed before the Control-space. If you wanted "a-breve" (not an IPA symbol, but comparative linguists use it anyway so we stuck it in the table) you'd type "c" to select it.

For IPA modifiers and some other characters, you type certain characters outside the a-z range before you type Control-space:

- Glottal stop and similar characters: type "**?**" and Control-space.
- Symbols for clicks: type "**!**" and Control-space.
- Modifier characters: type "^" (Shift-6) and Control-space for some, "**:**" (colon) and Control-space for others.
- Prosodic modifier characters: type "**:**" (colon) and Control-space.

For IPA diacritics that go under, above or through the base character proper:

- Diacritics that mainly go under the base character, type "**,**" (comma) and Control-space.
- Diacritics that go above the base character, type "~" (tilde) and Control-space.
- Diacritics that mainly go through the base character, type "-" (tilde) and Control-space.

Or, on a single line: **Ctrl-Sp after ? glottal; ! click; ^ : modifier; , ~ - diacritic**

Short Name	Datum
Mak	l e k o ʔ
Knj	r a u ŋ
Sel	t a h a
BuB	d a u ŋ
BuL	d a u ŋ
Cam	d a u̯ n
MBa	d a u̯ n
MMj	d a u̯ n
MPa	d a u̯ n
MSe	d a u̯ n
MMl	d a u̯ n
Mmu	d a u̯ n

IPA transcriptions from the Wordcorr Data

panel. Diacritical marks are not well aligned here, but they are elsewhere.

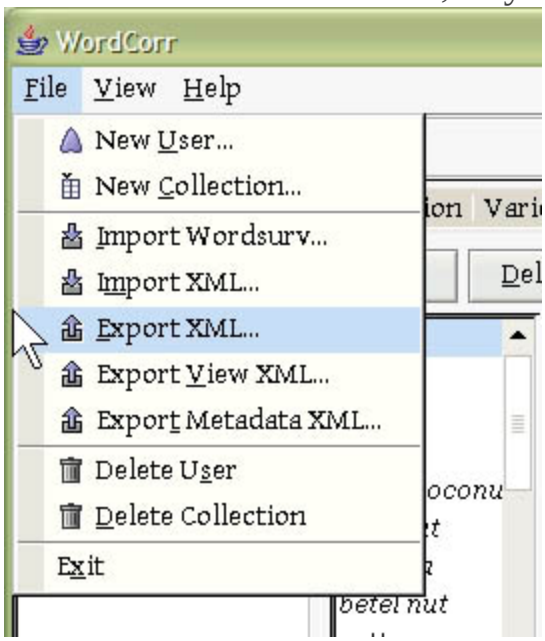
Note: When you're typing in data, you don't use all the characters in the International Phonetic Alphabet. For the ones you use a lot, your brain quickly learns the motor sequence (like "a" Control-space "2" for ash). If you touch type, you'll soon find yourself touch typing the part of IPA you use without a blink. You don't even have to take your hands off the keyboard to look for a mouse or a reference card. Install Wordcorr and give it a try.

Note: If you'd like to use Easy IPA in an application of your own, the Open Source Java source code is on the SourceForge CVS (anonymous access OK) as gui/input/IPAKeyListener.java. Please give credit to the National Science Foundation and Wordcorr.

Exporting for backup

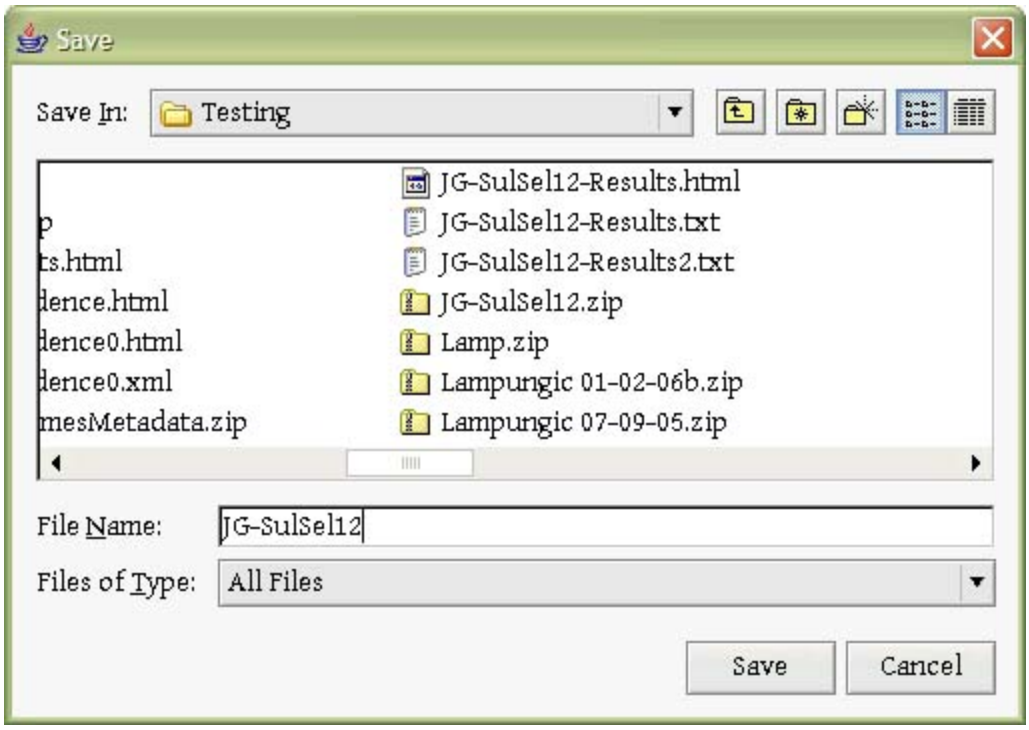
We've made it totally easy for you to back up your data and analyses at any time. The reason is that although every linguist in the world recognizes the need for backups, if you get personal and ask them about their own backups, they'll – well, you know what they'll say.

Step One is to click on the File | Export XML menu in the upper left part of the main Wordcorr window, as you see below in Figure 1.



Wordcorr's File | Export XML menu item selected.

Step Two is to choose the directory you want to save the backup file in, then put a file name into File Name for the backup of the collection you are working on. Figure 2 shows the Testing folder in Save In and a file name in File Name. Wordcorr will add a .zip extension to it.



The Export XML dialog box with a file name assigned to the backup.

- The file name you type into File Name has no dot and extension (like .xml or .zip) on the end. Let Wordcorr do it.
- The finished file will appear in the directory you designated for Save In.
- How long it takes depends on how many entries, varieties, and annotations you have, and how fast your computer is.

Back up your active Wordcorr project every day or so. You can use the same file name every time, as you see in Figure 2, or you can append a date to it in File Name. If you use a format like filename-2006-08-09, the backups will sort themselves into date order. But you may not want to keep that many around. (If you do, contact us at mail.wordcorr.org.)

Whenever you have access to the Internet (in the field you may be beyond the reach of even telephones), email your latest backup file to two or three friends you trust to keep it safe for you. If they are on different continents,

so much the safer. Also make arrangements for your university or office library to receive a copy for safekeeping. If you have a relationship with a permanent, professsionally curated archive, by all means email a copy to them.

If you can burn it to a CD or DVD, that's a good idea too.

Note: The same file you use as backup to safeguard your data and analyses also works for dialoguing with your colleagues, or for passing around homework assignments for your students. We'll get into that aspect of the scholarly community later.

Exporting part of a collection

You will probably use Export XML (from the File menu) every day to back up your work. Exported files also serve as your main mode of communicating what you are doing to your colleagues. They contain

- Your personal metadata (essentials only)
- Metadata about the collection
- Metadata for each speech variety
- The entries that contain the raw data
- Each analytical view you have created, with its annotations and tabulation results

But you may want to send only one of your analytical views along with the metadata and data. Export View XML, also on the File menu, lets you do that. It exports the view you are in when you invoke it, and ignores other views (except the Original view, because every collection has to have an Original view or Wordcorr won't work right).

Note: When you give a one-view export its file name, make sure it's NOT the same name as the complete collection! That could get you into trouble if you have an equipment failure and have to restore the collection from its current backup file. Instead of JG-SulSel for a one-view export, use a distinctive name like JG-SulSel-OneView or append the name of the specific view, like JG-SulSel-Reanalyze.

There's a related Export command, Export Metadata XML. It's intended to pass along just the metadata for you, your collection, and the varieties in the collection. But it also includes the data and some other things.

So it's best to ignore it until we clarify what we really need to do with it. Sorry for now.

Restoring from backup

The previous module showed you how to make a backup file. It suggested that you always keep an electronic copy in a safe place, preferably on another continent.

The acid test? A tsunami hits your field location and sweeps your house and your laptop out to sea. Fortunately, you were out of the way at the time. (No kidding; it happened to a friend of mine in pre-Wordcorr days.)

Eventually you get another computer, another house, ... From there, the steps are

1. Download and install the Wordcorr program. (If it's a new computer, you'll want to choose the First Installation option rather than the Upgrade option, since there's nothing in the new computer that you can upgrade.)
2. Have whoever keeps your backups send you the latest one as an email attachment.
3. Save it to your file system.
4. Start Wordcorr, select the File menu, select Import XML.
5. Select the directory where you saved your backup file, then select the file.
6. Click the "Open" button.

Your collection is restored to Wordcorr in a few minutes.

Note:I wish keeping backups for the rest of your computer were as straightforward. If you've kept your incremental backups in the same room as your computer, a Category Five disaster would wipe them out with it. But they might survive a lesser disaster.

Import from XML

You've already seen that Wordcorr saves everything in the form of XML files. An XML file has a tightly designed structure that keeps metadata, data, and the results of analysis sorted out at all times. A Wordcorr backup file is an XML file. A file for sharing a Wordcorr connection with another scholar or a student is the same thing, an XML file that preserves the structure of everything in a Wordcorr collection. A file for archiving is the same XML file, because it can be adapted to new media and new concepts years into the future.

So whenever you export an XML collection file, you have something that you or a colleague can import into the same computer to restore your work after a malfunction, or into another computer so that another pair of eyes can look at the same data and analysis.

The file you get will not be labeled as an XML file. It will be automatically compressed and presented as a ZIP file. But what is compressed is the XML file. Whether you refer to it as JG-Mind5.XML or JG-Mind5.ZIP, Wordcorr will check to see if it is compressed, and if so, it will extract the XML file from the ZIP file. If it isn't compressed (though it usually will be), Wordcorr knows enough to not try to uncompress it.

Note: The reason for compressing everything coming out of Wordcorr and uncompressing everything coming into Wordcorr is simple and humane. Most comparative linguists are occasionally out in the field, either without regular access to the Internet or at best with very slow access. That's where the interesting languages are. For them, the smaller their files are, the better.

That makes it easy to back up your work at least once a day. Comparative phonology is complicated enough that you don't want to have to backtrack because of a slipup or a computer crash.

Being able to import Wordcorr files also makes it easy for colleagues to work on large data sets as a team -- though you'll need caution if two people are working on the same part of the same view.

Note: If there is a chance two people's work on the same collection might get snarled up, the first precaution to take is to have your information technology people set you up with a versioning repository like CVS or SVN. See them for details.

XML is not the only way to take in data, and there are various ways to output data besides XML. The next section sketches some of these. But XML is the workhorse, safe and useful for all the main transport of information.

Import from Elsewhere

Many linguists have already put data into a computer. They may have had no access at the time to any standardized scheme for showing the structure of their data set. But at present there are two paths for getting data into Wordcorr from another computer readable form: spreadsheets and WordSurv. People will probably invent more paths as they are needed.

If you have comparative word lists in a spreadsheet, chances are good that you can copy them en masse into a special spreadsheet constructed by Maria Faehndrich, which a Python script then transforms into a Wordcorr XML file. The special spreadsheet, the Python script, and complete instructions are in the "utilities" package on Wordcorr's download site.

The idea is this: word lists are usually displayed together with the names or codes of speech varieties going with across the top of a table, and the meanings of the forms down the left side of the table. The forms themselves fill in the boxes.

You copy your word lists in a block from your spreadsheet to the place indicated on the special spreadsheet. You also fill in all the metadata for your collection on the special spreadsheet. Then you save that spreadsheet as a tab-delimited file (you may need to get a computer guru to show you how) and aim the Python script at the result (the same guru can probably handle it). You import the resulting file into Wordcorr.

A program called WordSurv has been in use for some time. Its original aim was to provide lexical similarity counts for possibly related speech varieties. It continues to be used as a way of storing word lists in a computer, even though few linguists give credence to its conclusions about relationships because they have been shown to be unreliable in the statistical sense.

Wordcorr makes use of WordSurv's very sensible plan for data storage, but applies widely tested and agreed upon best practices to produce its results. Wordcorr contains a converter for the first versions of Wordsurv, with instructions in Wordcorr Help. More recent versions can be converted by the Palmsurv converter.

WordSurv and older spreadsheets were unable to handle phonetic symbols, so linguists used their imagination: N for angma, E for epsilon, 7 for glottal stop, on and on and usually different for each collection. One of the spreadsheet utilities includes a table in which you can put the Unicode characters of the International Phonetic Alphabet that you want to convert to. For imports from WordSurv, however, hand editing on the Data panel seems to be the only feasible solution so far.

What We've Covered

In the preceding sections you've acquired a vocabulary for talking about the elements of comparative phonology as manifested in Wordcorr -- Wordcorrspeak, we might call it. You have the concepts that define the framework in which Wordcorr operates:

- User
- Metadata
- Collection
- Variety
- Datum
- Easy IPA
- View
- Backup
- Export

What you don't have yet are the operations at the heart of doing comparative phonology:

- Annotate
- Tabulate
- Refine
- Share

There's a reason for not going into them in a tutorial like this one. It's that they are best understood by loading up Wordcorr (details in the Installation tutorial), going to the Wordcorr Web site (wordcorr.org), and trying it yourself, using both the Web site and the Help facility that downloads with Wordcorr itself.

What We Haven't Covered

From here on you don't want a step-by-step tutorial. You want to be able to jump around via hyperlinks. That's how the Wordcorr Web page is set up.

Take a look at the home page first. Click on the screen shot of a fragment of the final product. That will show you where you are headed.

From the home page, click on "New user". The first part of the page is what is covered in this tutorial. The second part is your guide to the various operations you will be learning to perform. Work out from this page in all directions, and you'll have a pretty complete idea of how Wordcorr works!

We'd like to hear from you, whether everything goes smoothly for you or you hit rough spots. Especially the rough spots, because maybe then we can fix them. Use the feedback facility at the bottom of every page in this tutorial.

Set up Wordcorr

Install Wordcorr. Wordcorr is free. Wordcorr runs on any Windows system from Windows 98 on. (We're working on Macintosh and Linux versions, but they're not ready yet.) Choose the type of download Wordcorr you need, and follow the instructions. Then come back to this page for information on how to run Wordcorr.

- First-time installation on Windows
- Upgrade an earlier version already installed on Windows

Run Wordcorr. Once you have Windows on your computer, you can start it up any time. You can have the installer program put a Wordcorr icon on your desktop or one your task bar near the Start button. Or you can reach it through the Start button itself.

Set up your Wordcorr identifiers. Once Wordcorr is running, it needs a small amount of public information about you before you can start to actually do anything linguistic. Click here to find out about user metadata.

First-time installation

You have a laptop or desktop computer that uses the Windows operating system, any one from Windows 98 to NT to 2000 to XP. If you're reading this, you must know how to start your computer already. And if you're reading this, you must know how to connect to the Internet.

If you have never installed Wordcorr on this computer, you're on the right page to do a first-time installation. If there is already a copy of Wordcorr on this computer, you should go to the Upgrade page to install the current version. Or if you bring up the version of Wordcorr that's already there, click on the Help menu item and the About Wordcorr item, and it tells you that it is Version 2.0.0, then you don't need to do anything.

What kind of Internet connection are you using? Wordcorr is a fair sized program, 20.1 megabytes. It can take a long time to download if all you have is a dial-up telephone connection through a modem, more than 48 minutes at the fastest phone modem speed, and over 3 hours at the slowest.

If you have any way of connecting to the Internet through a broadband connection, use that for the installation. Once Wordcorr is installed, however, you can use it without broadband. Broadband connections come through cable television companies, a DSL link provided by your telephone company, or a direct arrangement often found in universities and corporate offices. Internet cafes sometimes don't have broadband; many Starbucks and Borders, airports and shopping malls, do.

Note: If you can't work out a broadband arrangement, find someone who has one. Ask them to download Wordcorr for you, but instead of installing it on their computer, transfer it to a CD-ROM so that you can install it from there. (If you can't read a CD-ROM, your computer's capacity may be too small to handle the 20.1 megabytes that make up Wordcorr; maybe it's time to change computers.)

When you click on Download at the very end of this module, it takes you to the place on the Internet where you can always find the latest version of the program: a Web site called SourceForge.net that is the distribution point for millions of downloads. SourceForge confirms that you are downloading Release 2.0.0 for initial installation, and shows you a list of **mirror** Web sites around the world that you can download from. Pick one reasonably near you and click on the "Download" link to its right.

A "File Download" box will come up and ask if you want to open the file being downloaded, or save it. Choose "Open".

Note: If someone else is downloading for you, tell them to choose "Save". That will put the file on their disk, where they can put it on a CD-ROM for you. To load it from a CD-ROM, bring up the directory that contains wordcorr-2-0-0-install.exe and double click on the file name.

From that point on, you're in familiar territory. The downloaded file contains an installer, which offers to put Wordcorr in C:\Program Files\Wordcorr for you. Accept the offer, and any others that it makes. It also puts some files with internal configuration information in a special folder, and sets up a copy of the Java Runtime Environment where it knows how to find it -- behind the scenes modules that you don't need to worry about.

Got all that? Good! Now here goes: Download

Upgrading Wordcorr

You have a laptop or desktop computer that uses the Windows operating system, any one from Windows 98 to NT to 2000 to XP. You already have Wordcorr installed on your computer, but now it is time to upgrade it to a new release. (If you **don't** have Wordcorr installed, click here to find out how to install it the first time.)

Make sure you have current backup files exported for every collection that is already in Wordcorr on your computer. Then go ahead with the upgrade.

The current Wordcorr upgrade (2.0.0) applies to any Wordcorr release 1.3.7 or after.

Note: For earlier releases, the internal database structure is a little different. Make sure you have all your collections backed up. Then do Start, Search, Files and Folders, and search for ".wordcorr" (that's DOTwordcorr, and leave off the quotation marks). Click on that folder to open it, and delete the files in it, but don't delete the folder. Then when you finish the upgrade, Wordcorr will recreate its internal database in the correct form. For releases 1.3.7 or later, you don't need to clear the .wordcorr folder.

What kind of Internet connection are you using? The Wordcorr upgrade is 7.1 megabytes, smaller than the full installation. It can still take a long time to download if all you have is a dial-up telephone connection through a modem.

If you have any way of connecting to the Internet through a broadband connection, use that for the upgrade. Once Wordcorr is upgraded, however, you can use it without broadband. Broadband connections come through cable television companies, a DSL link provided by your telephone company, or a direct arrangement often found in universities and corporate offices. Internet cafes don't usually have broadband.

Note: If you can't work out a broadband arrangement, find someone who has one. Ask them to download the Wordcorr upgrade for you, but instead of installing it on their computer, transfer it to a CD-ROM so that you can install it from there. (If you can't read a CD-ROM, your computer's capacity may be too small to handle Wordcorr; maybe it's time to change computers.)

When you click on Download at the very end of this module, it takes you to the place on the Internet where you can always find the latest version of the upgrade. It's a Web site called SourceForge.net that is the distribution point for millions of downloads. SourceForge confirms that you are downloading Release 2.0.0 for upgrade, and shows you a list of **mirror** Web sites around the world that you can download from. Pick one reasonably near you and click on the "Download" link to its right.

A "File Download" box will come up and ask if you want to open the file being downloaded, or save it. Choose "Open".

Note: If someone else is downloading for you, tell them to choose "Save". That will put the file on their disk, where they can put it on a CD-ROM for you. To load it from a CD-ROM, bring up the directory that contains wordcorr-2-0-0-upgrade.exe and double click on the file name.

From that point on, you're on familiar territory. The file contains an installer, which offers to put Wordcorr in C:\Program Files\Wordcorr for you. Accept the offer, and any others that it makes. It also puts some files with internal configuration information in the special .wordcorr folder if you have deleted the ones that were there. It does not set up a new copy of the Java Runtime Environment, because it knows how to find the one you already have.

Got all that? Good! Now here goes: Download